scientific reports

OPEN

A coarse-to-fine registration method for multimodal retinal images

Jianguo Xu[1]✉, Sen Zhang[1], Jianxin Shen[1], Jiahao Wei[1], Sukun Tian[2], Jin Yao[3], Zhipeng Yan[3] & Fen Zhou[3]✉

Manual preoperative image registration for central serous chorioretinopathy (CSCR) is labor-intensive and irreproducible. While rigid registration robustly aligns images globally, it misses fine details. Non-rigid registration, though excellent for local refinement, performs poorly with large discrepancies. Therefore, this study presents a coarse-to-fine registration method for multimodal retinal images to address the aforementioned issues. First, a three-step coarse registration strategy is designed that integrates keypoint pair detection and matching via a YOLOv8-pose network, further optimizes keypoints through a post-processing technique, and achieves initial alignment via affine transformation. On this basis, a dual-component fine registration strategy is then implemented, where disentanglement learning eliminates modality-specific variations while preserving essential vessel structures required for registration, and deformable network generates optimized deformation field to refine the coarse alignment locally, ultimately enabling high-precision image registration. Comprehensive qualitative and quantitative experiments were conducted on the CSCR clinical dataset, which includes both color fundus (CF) and fundus fluorescence angiography (FFA) images, to evaluate the proposed method. With Dice and Dice$_s$ scores of 0.6759 and 0.4977, the method performs comparably to existing approaches, suggesting its potential application value for CSCR preoperative planning.

Keywords CSCR, Coarse registration, Fine registration, Disentangled network, Deformable network

Ocular diseases represent one of the significant threats to human health. Ophthalmologists employ common treatment approaches including pharmacological interventions and surgical procedures, such as using medications to treat uveitis[1,2] and applying laser therapy for CSCR. As widely recognized, retinal laser surgery is an effective intervention for CSCR, promoting the absorption of subretinal fluid through precise photocoagulation of leakage points. The technique, with its minimally invasive nature, ease of operation, and short recovery period, has become a routine clinical choice for treating CSCR. Retinal image registration in preoperative planning is crucial for ensuring the precision of laser surgery. Currently, in clinical ophthalmology, doctors mainly rely on manual alignment of multimodal retinal images, which is labor-intensive and has poor reproducibility, and the results are affected by subjective factors, making it difficult to control registration accuracy. Therefore, automatic registration of multimodal retinal images has become an important direction for exploration.

Rigid registration is a fundamental method in medical image registration and mainly includes intensity-based registration and feature-based registration. Legg[3] employed mutual information as a similarity metric to achieve registration between retinal color photographs and scanning laser ophthalmoscope images. Reel[4] investigated the application of the Expectation Maximization for Principal Component Analysis based Mutual Information algorithm in retinal image registration, which combines spatial information with mutual information to effectively boost registration performance. Additionally, Lange[5] proposed a Normalized Gradient Fields distance measure to handle the registration of two-dimensional and three-dimensional CT images. Yang[6] developed the GDB-ICP algorithm capable of processing image pairs exhibiting low overlap, significant scale variations,

[1]College of Mechanical and Electrical Engineering, Nanjing University of Aeronautics and Astronautics, No.29 Yudao Road, Qinhuai District, Nanjing 210016, People's Republic of China. [2]Center of Digital Dentistry, Peking University School and Hospital of Stomatology & National Engineering Research Center of Oral Biomaterials and Digital Medical Devices & NHC Key Laboratory of Digital Stomatology, No.22 Zhongguancun South Street, Haidian District, Beijing 100081, People's Republic of China. [3]The Affiliated Eye Hospital of Nanjing Medical University, No.138 Hanzhong Road, Gulou District, Nanjing, People's Republic of China. ✉email: xu_nuaa_edu@hotmail.com; zf520simple@163.com

and illumination discrepancies, successfully achieving cross-domain registration for both natural and medical images. Ghassabi[7] introduced the UR-SIFT-PIIFD algorithm to address the standard SIFT's limitation in feature point extraction for multimodal retinal images, significantly improving feature point distribution uniformity. Wang[8] proposed a scheme called SURF-PIIFD-RPM that combines a SURF detector for PIIFD descriptor extraction with Robust Point Matching for outlier rejection, demonstrating particularly strong performance in low-overlap regions. Chen[9] introduced a new registration method that captures the key geometric characteristics of vessel bifurcations and embeds this information into feature vectors to assess similarity between images and achieve registration. Wang[10] designed a deep learning-based three-stage framework for multimodal retinal image registration, employing segmentation, feature detection and description, and outlier rejection networks to align color photographs with infrared images. Subsequent work[11] enhanced this framework through phase map integration in the segmentation network, enabling registration of color photographs with both angiography and infrared images. The aforementioned approaches exhibit remarkable efficacy in rigid image registration, constituting pivotal research directions in medical image analysis.

In recent years, the rapid development of deep learning has demonstrated outstanding performance in industrial scenarios, such as cyber threat detection in maritime contexts[12], as well as remarkable capabilities in various medical imaging tasks including disease diagnosis[13–16], lesion segmentation[17–21], and medical image super-resolution[22]. It is also increasingly showcasing its advantages in non-rigid registration. Non-rigid registration serves as another critical paradigm in medical image registration, addressing complex deformations (e.g., stretching, compression, and twisting). Lee[23] utilized a CNN to detect vessel bifurcation points in retinal images, enabling geometric transformation-based registration between CF and OCT modalities. Zhang[24] raised a neural network-based retinal image registration pipeline that jointly performs vessel segmentation and deformable registration. This approach employs style loss to transform retinal images of different modalities into a consistent representation, achieving non-rigid registration through a deformable registration network. The research team[25] also innovatively put forward a two-step registration approach, utilizing rigid registration for coarse alignment of multimodal images, followed by refinement of the registration results through a deformable network. Inspired by the VoxelMorph network, Martínez[26] proposed a weakly supervised deep-learning framework for deformable registration of FFA and OCTA images. Santarossa[27] proposed MedRegNet, a lightweight descriptor module compatible with feature-based pipelines, which can improve the robustness and registration performance of classic detectors. The above non-rigid registration methods demonstrate respective advantages and collectively serve as a key driver for advancing the development of image registration.

However, it should be noted that rigid registration performs exceptionally well with multi-modal retinal images under large-discrepancy conditions, yet shows limitations in local alignment. In contrast, non-rigid registration excels at fine local registration but proves less effective with significant discrepancies. Thus, we cannot help but wonder whether these two image registration schemes could be integrated to capitalize on their respective strengths while mitigating their weaknesses, thereby better enabling automated registration for CSCR multimodal retinal images. This motivation has led us to propose a coarse-to-fine multimodal retinal image registration method, comprising a three-step coarse registration strategy and a dual-component fine registration strategy. The main contributions are as follows:

- First, this study establishes two datasets specifically for CSCR multimodal retinal image registration, providing a foundation for subsequent research.
- Second, we propose a three-step coarse registration strategy that initially integrates keypoint detection and matching using the YOLOv8-pose network, subsequently optimizes keypoints through a post-processing technique, and ultimately achieves initial multimodal retinal alignment via affine transformation.
- Third, a dual-component fine registration strategy is developed, in which modality discrepancies are first suppressed while essential vessel structures are preserved through disentanglement learning, and deformation fields are subsequently generated through a deformable network to further refine the coarsely-registered images, yielding the optimally registered output.
- Finally, qualitative and quantitative experiments are conducted on the CSCR dataset to evaluate the efficacy of the proposed multimodal retinal image registration method.

The structure of the remaining part is as follows. Section "Related works" describes the related works. Section "Materials and proposed method" describes the materials and explains the implementation details of our proposed method. Section "Experiments and discussions" shows the experiments. Section "Limitations and future work" shows the limitations and future work. Section "Conclusions" concludes the research work.

Related works

Object detection

Object detection is a key research area in computer vision, entailing both object localization and classification[28]. The task involves detecting and classifying a varying number of objects in an image, such as the recognition of lesion[29] and abnormality[30]. Meanwhile, it also demonstrates excellent applications in areas such as high-quality detection of spike firings from hundreds of neurons[31] and cyber attack detection in shipboard microgrids system[32,33]. It is worth mentioning that YOLO, a representative single-stage deep learning detection network, has been widely used in various practical applications. Jian[34] proposed an improved YOLOv7-tiny algorithm, which is suitable for detecting occluded pedestrians in autonomous driving scenarios. Qian[35] enhanced YOLOv5s to achieve object detection for lightweight ships. He[36] applied the YOLO model to train and detect ground object targets in high-resolution remote sensing images. Hou[37] developed a rapid detection method for counting wheat seedling leaves in complex field scenarios based on an improved YOLOv8. Mugahed[38] applied YOLO to accurately detect the masses from the entire mammograms. These established applications substantiate YOLO's adaptability across multiple object detection domains. Nevertheless, its efficacy in detecting keypoint pairs within retinal images has yet to be investigated, providing the primary impetus for our current research.

Disentanglement learning

Disentanglement learning has demonstrated superior performance in image-to-image translation. Representative frameworks like MUNIT[39] and DRIT[40] employ disentanglement learning by factorizing images into two latent spaces: a domain-invariant content space and a domain-specific style space. This paradigm has not only advanced translation performance but has also been successfully applied in medical image processing. To address class imbalance in fatty liver disease classification, Huang[41] proposed the ICFDNet network, significantly improving both overall accuracy and inter-class performance balance. Jin[42] presented a CCNet, a decoupling network that decomposes low-light image enhancement into brightness enhancement and colorization subtasks for customized results. Qin[43] proposed the UMDIR for unsupervised deformable registration of multimodal brain images via disentangled representations. Liao[44] developed the ADN network, effectively addressing the issue of metal artifacts in medical imaging caused by implanted metal objects in patients. These studies highlight the effectiveness of disentanglement learning in medical image tasks.

Deformable registration

Deformable registration is an image processing technique primarily used to align two or more images. It establishes non-rigid transformation relationships between images, enabling them to match each other in terms of shape, size, and position. This technique is widely applied in medical image processing, remote sensing image analysis, and other fields, where it effectively handles complex deformations in images to achieve precise image fusion, comparison, or analysis. He[45] proposed a 3D deformable registration algorithm for dose tracking and optimization in targeted radiotherapy for prostate cancer. Mohamed[46] introduced a method to deformably register 3D brain tumor images to a normal brain atlas. Michael[47] employed a deformable image registration based on a biomechanical model to warp expiratory CT images to inspiratory CT images, thereby calculating dose accumulation over the entire respiratory cycle. Perez-Rovira[48] developed an inter-frame deformable registration algorithm for ultra-wide field view retinal fluorescein angiography sequences, addressing the critical challenge of aligning temporal FA frames with evolving retinal vessel structures.

It can be observed that object detection, disentanglement learning, and deformable registration are widely distributed across various visual tasks. Although these approaches differ in their specific application targets, the coupling idea of relevant skills and practical events have greatly inspired the construction of the solutions in this paper.

Materials and proposed method

Materials

The CF and FFA images of CSCR used in this study were provided by the Affiliated Eye Hospital of Nanjing Medical University, with approval from the hospital's Medical Ethics Committee. The age of CSCR patients is predominantly distributed between 20 and 50 years, with a higher prevalence among young and middle-aged males. The images are captured using a fundus camera and include both acute and chronic types of CSCR. The study adhered to the principles of the Declaration of Helsinki. The CSCR dataset for this task consists of Dataset-1 and Dataset-2, with the original collection containing a total of 306 pairs of CF and FFA images.

- Dataset-1: Comprising 216 pairs of CF and FFA images. The dataset is divided into 138 pairs for training, 9 pairs for validation, and 69 pairs for testing. As shown in Fig. 1, a data augmentation approach (including horizontal mirroring, clockwise rotation by 5°, and counterclockwise rotation by 5°) is first applied to expand the training set to 828 images and the validation set to 54 images. Subsequently, image stitching and keypoint annotation are performed. Image stitching refers to the process of horizontally connecting CF and FFA images without overlap to form a complete composite image. Keypoint annotation involves marking vessel bifurcation points following a dispersed distribution principle on both modality regions of the stitched image, drawing bounding boxes that encompass the bifurcation points, and assigning label categories to correspond-

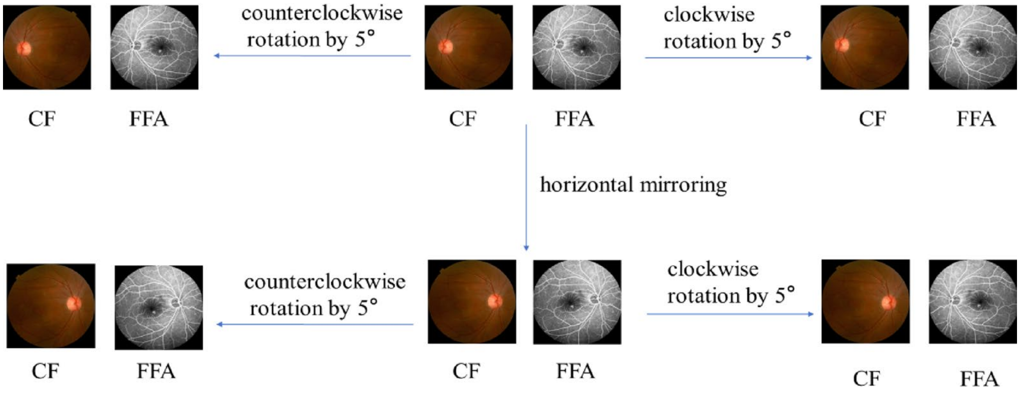


Fig. 1. Data augmentation.

ing points. These operations are carried out jointly by ophthalmologists and researchers through annotation software, guided by the practical image registration expertise of the ophthalmologists. After image stitching and keypoint annotation, this dataset is used to train and validate the YOLOv8-pose network in the coarse registration stage. Besides, 147 pairs of CF and FFA images are adopted to train the deformable network in the fine registration stage.

- Dataset-2: Comprising 90 pairs of CF and FFA images. Through the same data augmentation operations (Fig. 1), the dataset is expanded to 540 image pairs. These augmented CF and FFA images are then combined with pseudo labels (i.e., the binary vessel labels from the DRIVE[49] dataset for training the disentangled network in the fine registration stage. It should be noted that the fine details at the vessel tips in the pseudo labels were erased, as preliminary experiments revealed that these delicate structures could introduce minor artifacts in the vessel maps produced by the disentangled network. Performance evaluation in this stage continues to employ the 69 pairs of CF and FFA images from Dataset-1, ensuring consistent benchmarking throughout the study.

The proposed method

This section describes the implementation details of the proposed coarse-to-fine registration method for CSCR multimodal retinal images. The method is decomposed into two subtasks with corresponding processing stages. As shown in Fig. 2, the coarse registration stage executes a three-step registration strategy involving keypoint pair detection and matching, keypoint optimization, and initial registration through affine transformation. The fine registration stage adopts a dual-component fine registration strategy, where a disentangled network first removes modality-specific characteristics from CF and FFA images while maintaining the essential vessel structures, after which a deformable network generates a deformation field to further adjust the coarsely-registered images, yielding the final registration outcome.

Coarse registration stage

In the coarse registration stage, this study designed a three-step coarse registration strategy for the case of large discrepancies between CF and FFA images, which consists of keypoint pair detection and matching, keypoint optimization, and initial alignment based on affine transformation.

- Keypoint pair detection and matching
  In traditional rigid registration schemes, keypoint detection and matching are independent processes, resulting in cumbersome procedures. Here, leveraging the previously created CSCR Dataset-1 with stitching attributes and keypoint correspondences, along with the Yolov8-pose network, we unify these two processes into a single task, thereby simplifying the registration pipeline. The implementation process is illustrated in Fig. 3. First, the training images are fed into the YOLOv8-pose network for end-to-end training. Subsequently, the trained model performs inference on the testing images. The model's output includes two types of information: (1) visual detection results, consisting of annotated testing images with keypoint pairs, bounding boxes, and their corresponding categories and confidence scores; and (2) textual detection data, which records the category and coordinate information of keypoint pairs, and bounding box coordinates in detail. For additional details of the YOLOv8-pose network, please refer to Reference[50]. It is important to note that in Dataset-1, the same vessel bifurcation points in the CF and FFA images have been annotated with predefined category labels and bounding boxes. These pre-defined labels and images guide the YOLOv8-pose network to update its parameters, equipping it with the capability to automatically identify and match the homologous vessel bifurcation points across the two modalities. Obviously, the model's training is not limited by the field-of-view differences between CF and FFA images. Consequently, the trained model can successfully detect and match these point pairs whenever the same vessel bifurcation points are present in both images, demonstrating robust performance even under significant disparity conditions.

- Keypoint optimization
  The detection performance of the YOLOv8-pose network for vessel bifurcations is closely related to image characteristics. When bifurcations are distinct and clear, the model can predict more keypoints with smaller localization errors. However, when bifurcations are unclear or sparse, both the prediction accuracy and the number of detected keypoints decrease significantly, adversely affecting subsequent registration results. Since vessel bifurcations represent regional features rather than single pixels, inconsistencies in manual labeling may further compromise prediction reliability. Moreover, as the affine transformation model requires at least three non-collinear keypoint pairs to compute the transformation matrix, careful selection of appropriate keypoint pairs from the YOLOv8-pose network's predictions warrants attention. To address these issues, we propose a simple post-processing technique for keypoint optimization, which consists of keypoint relocation and keypoint pair selection. The former involves grayscale conversion, B-COSFIRE filtering, skeletonization, and neighborhood search to facilitate the localization of bifurcations near the initial keypoints. Among these operations, the first few operations in keypoint relocation are common image processing techniques. Here, we elaborate on the implementation details of the neighborhood search. Specifically, for each initial keypoint predicted by YOLOv8-pose, a square region centered on it with a side length of 21 pixels is first delineated on the skeletonized vessel map. A 3-pixel square window then scans this region from the top-left to the bottom-right corner with a step size of 1 pixel, while counting the number of vessel pixels within each scanning window. The central pixel of each scanning window containing more than 3 vessel pixels is designated as a candidate point. The distances between these candidate points and the initial keypoint are calculated, with the

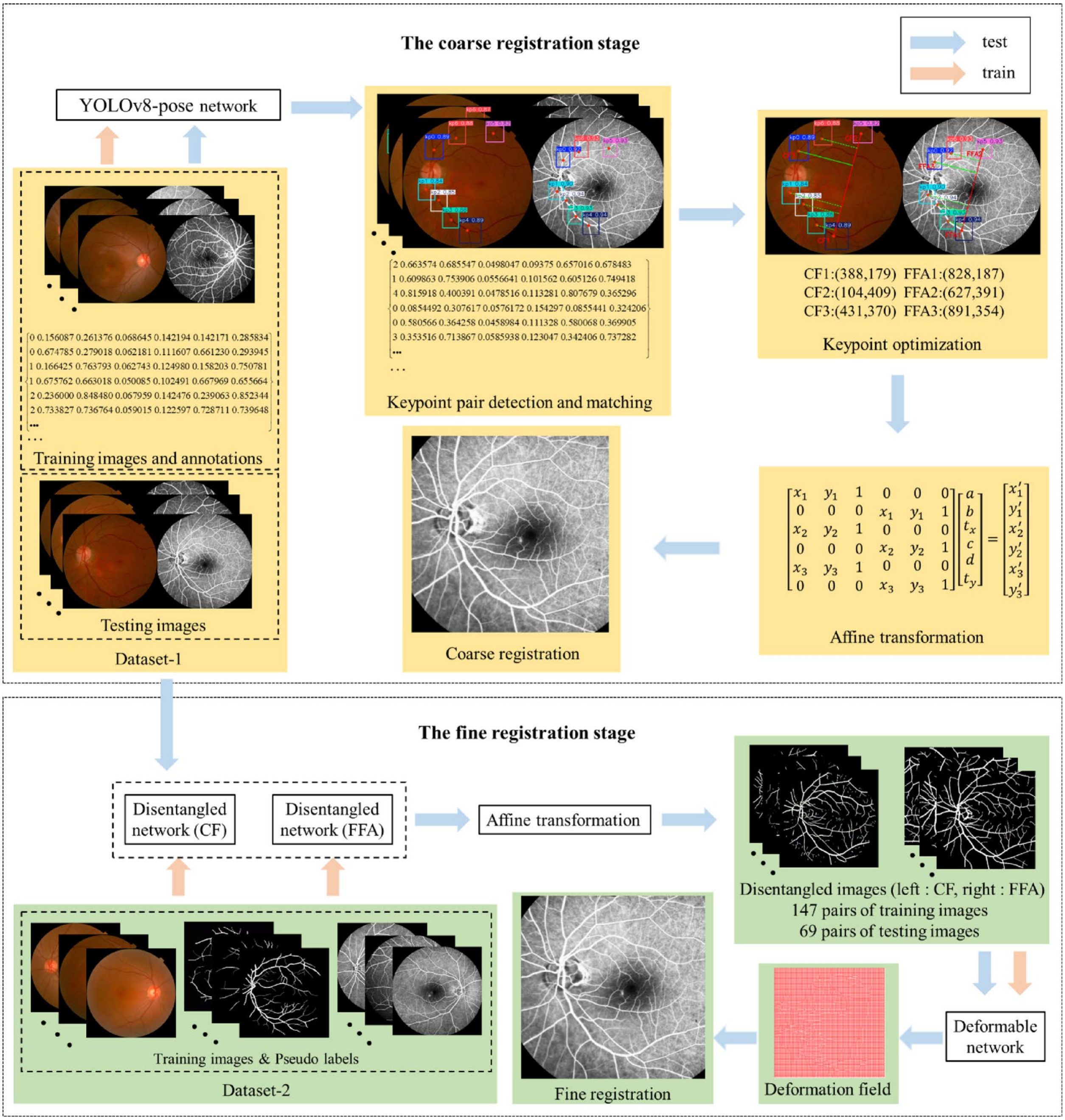
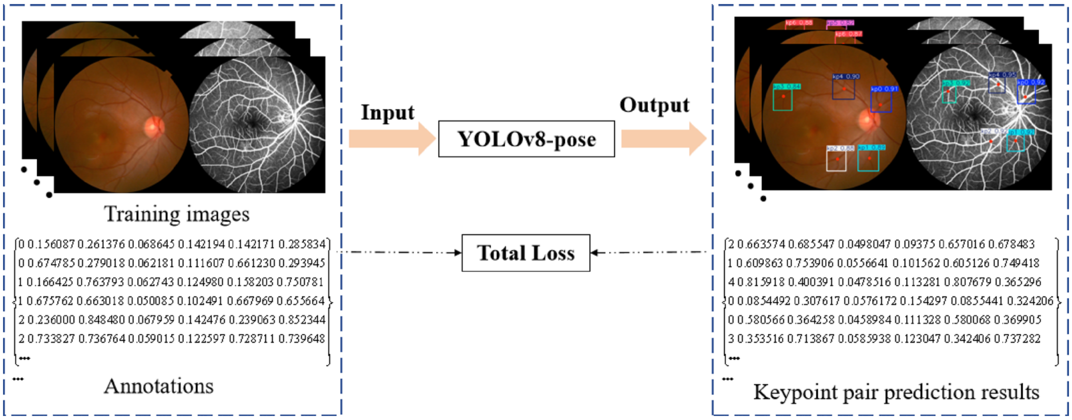


Fig. 2. The implementation details of the proposed coarse-to-fine registration method: (Upper) Coarse Registration for global alignment via keypoint pair detection and matching, followed by keypoint optimization and affine transformation; (Lower) Fine Registration using a disentangled network to preserve vessel structures, followed by a deformable network for local refinement.

candidate point corresponding to the smallest distance being selected as the relocated point (corresponding to each green point shown in the keypoint relocation section of Fig. 4). Following this operation, the keypoint pair selection step is subsequently executed. This step ensures better spatial dispersion of keypoint pairs in the image by maximizing (1) the distance between the first two keypoints and (2) the perpendicular distance from the third keypoint to the line connecting the first two points. Figure 4 illustrates the keypoint optimization process.

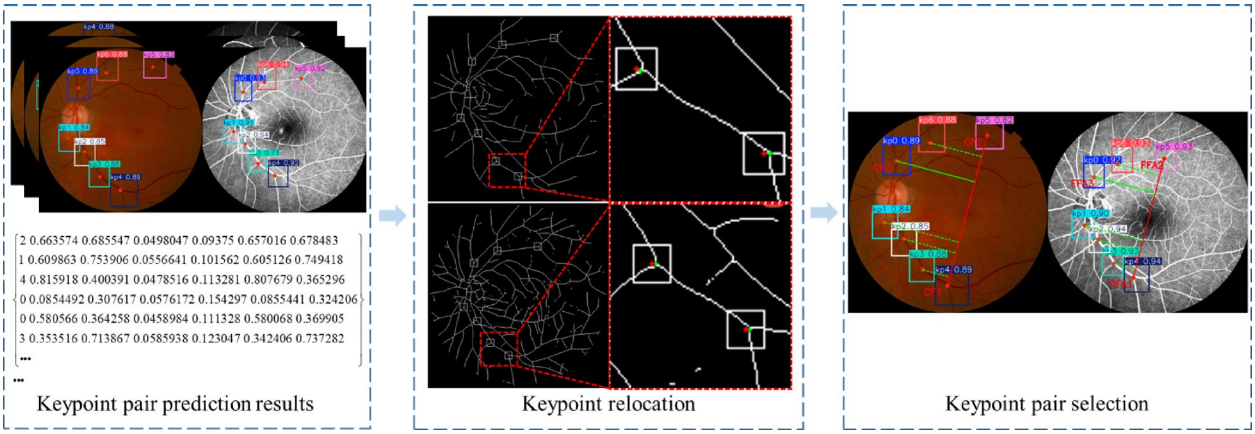
- Coarse registration
  After the keypoint optimization, an affine transformation is employed to achieve coarse registration of multimodal retinal images. As a generalized linear transformation model, the affine transformation not only incorporates all the properties of rigid transformations (including translation and rotation) but also enables scaling and shearing. Its core advantage lies in maintaining the parallelism of lines before and after the trans-



```
0 0.156087 0.261376 0.068645 0.142194 0.142171 0.285834
0 0.674785 0.279018 0.062181 0.111607 0.661230 0.293945
1 0.166425 0.763793 0.062743 0.124980 0.158203 0.750781
1 0.675762 0.663018 0.050085 0.102491 0.667969 0.655664
2 0.236000 0.848480 0.067959 0.142476 0.239063 0.852344
2 0.733827 0.736764 0.059015 0.122597 0.728711 0.739648
...
...
```

```
2 0.663574 0.685547 0.0498047 0.09375 0.657016 0.678483
1 0.609863 0.753906 0.0556641 0.101562 0.605126 0.749418
4 0.815918 0.400391 0.0478516 0.113281 0.807679 0.365296
0 0.0854492 0.307617 0.0576172 0.154297 0.0855441 0.324206
0 0.580566 0.364258 0.0458984 0.111328 0.580068 0.369905
3 0.353516 0.713867 0.0585938 0.123047 0.342406 0.737282
...
...
```

Fig. 3. Keypoint pair detection and matching based on YOLOv8-pose Network. The process yields two distinct outputs: (1) Visual detection results, displaying annotated images with keypoint pairs, bounding boxes, categories, and confidence scores; (2) Textual detection data, containing detailed records of keypoint pair coordinates and bounding box information.



```
2 0.663574 0.685547 0.0498047 0.09375 0.657016 0.678483
1 0.609863 0.753906 0.0556641 0.101562 0.605126 0.749418
4 0.815918 0.400391 0.0478516 0.113281 0.807679 0.365296
0 0.0854492 0.307617 0.0576172 0.154297 0.0855441 0.324206
0 0.580566 0.364258 0.0458984 0.111328 0.580068 0.369905
3 0.353516 0.713867 0.0585938 0.123047 0.342406 0.737282
...
...
```

Fig. 4. The keypoint optimization process : (1) Keypoint relocation refines initial positions through a sequence of grayscale conversion, B-COSFIRE filtering, skeletonization, and neighborhood search to accurately locate bifurcations; (2) Keypoint pair selection optimizes the spatial configuration by maximizing both the distance between the first two keypoints and the perpendicular distance from the third keypoint to the line connecting them.

formation, as well as allowing anisotropic scaling of images. These characteristics make it particularly suitable for medical image registration tasks. For instance, in retinal image registration, it can effectively correct scale differences and geometric deformations caused by different imaging devices or shooting angles. Compared with simple rigid transformations, the affine transformation can better handle complex spatial correspondences between multimodal images, thereby laying a solid foundation for subsequent fine registration. The mathematical formula for the affine transformation is as follows:

$$\begin{pmatrix} x_z' \\ y_z' \\ 1 \end{pmatrix} = \begin{pmatrix} a & b & t_x \\ c & d & t_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_z \\ y_z \\ 1 \end{pmatrix} \tag{1}$$

where (x_z, y_z) and (x_z', y_z') denote the coordinates before and after transformation, respectively. The linear transformation parameters a, b, c, and d govern geometric deformations including rotation, scaling, and shearing, while t_x and t_y represent the translation vector.

Fine registration stage

While coarse registration achieves initial alignment of multimodal retinal images, further refinement remains essential. To address this, we propose a dual-component fine registration strategy comprising: (1) a disentangled network that eliminates modality-specific characteristics while preserving the necessary vessel structures crucial for registration in both CF and FFA images, followed by (2) a deformable network that generates a deformation field to refine the coarse registration results.

- Modality information removal

As an outstanding work in disentanglement learning paradigms, ADN[44] has demonstrated promising performance in metal artifact removal. Therefore, we investigate its applicability to fine registration stage in our study. This paper decomposes multimodal retinal images into two complementary feature spaces based on a disentangled network: a content space shared across modalities and an attribute space specific to each modality. The content space focuses on extracting vessel structural features that are independent of the imaging modality, while the attribute space retains the unique imaging characteristics of each modality. This approach transforms the complex multimodal registration problem into a single-modality registration problem, significantly reducing the difficulty of registration. By separating modality-specific and shared features, it provides a more robust feature basis for subsequent deformable registration.

The structure of the disentangled network is shown in Fig. 5. It takes retinal images carrying modality information and retinal vessel images not carrying modality information as inputs, and outputs retinal images with modality information removed. The network comprises three encoders $E_I : I \to C$, $E^c_{I^m} : I_m \to C$, and $E^m_{I^m} : I^m \to M$, as well as two decoders $G_I : C \to I$ and $G_{I^m} : C \times M \to I^m$. The encoders are responsible for mapping images to the content space and modality space, while the decoders map the content space and modality space back to images. Specifically, given any two unpaired images $x^m \in I^m$ and $y \in I$, E_I and $E^C_{I^m}$ map x^m and y to the content spaces (c_x and c_y) respectively, while $E^m_{I^m}$ maps x_m to the modality space m. This process can be denoted as:

$$c_x = E^C_{I^m}\left(x^m\right) \tag{2}$$

$$m = E^m_{I^m}\left(x^m\right) \tag{3}$$

$$c_y = E_I\left(y\right) \tag{4}$$

where the decoder G_{I^m} maps the content spaces c_x and c_y along with the modality space m back to the images x'^m and y'^m, which can be denoted as:

$$x'^m = G_{I^m}\left(c_x, m\right) \tag{5}$$

$$y'^m = G_{I^m}\left(c_y, m\right) \tag{6}$$

where the decoder G_I maps the content spaces c_x and c_y back to the images x' and y' which can be denoted as:

$$x' = G_I\left(c_x\right) \tag{7}$$

$$y' = G_I\left(c_y\right) \tag{8}$$

where y'' is the result of encoding and then decoding the synthetic image y'^m, which should be able to reconstruct the original image, denoted as:

$$y'' = G_I\left(E^C_{I^m}\left(y'^m\right)\right) \tag{9}$$

The loss function of the disentangled network consists of five parts: two adversarial losses L^I_{adv} and $L^{I^m}_{adv}$, modality consistency loss L_{mod}, reconstruction loss L_{rec} and self-reduction loss L_{self}. These loss functions

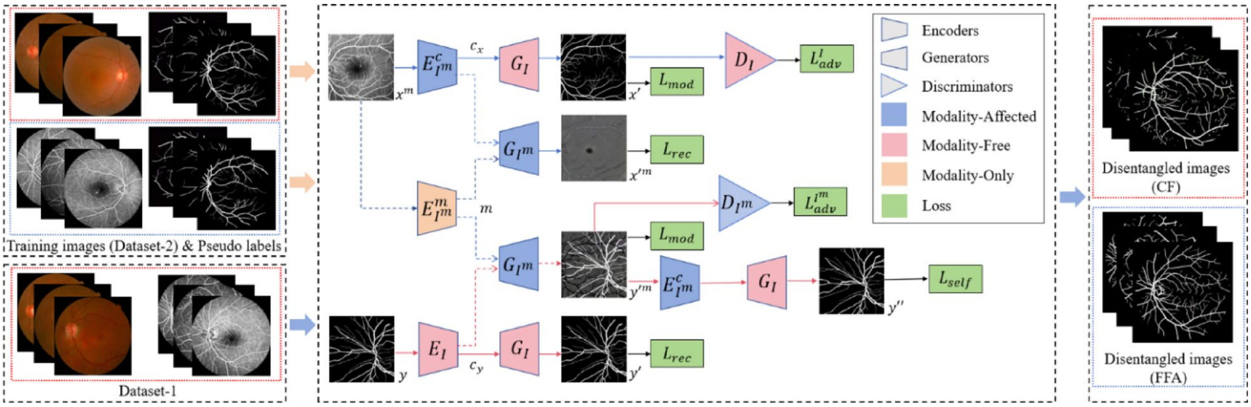


Fig. 5. The process of modality information removal.

collectively drive the disentangled network to remove modality-specific information from multimodal retinal images. L_{adv}^{I} and $L_{adv}^{I^m}$ are used to distinguish between generated and real images, thereby enhancing the quality of the generated images. L_{rec} ensures that the encoding and decoding processes can accurately restore the original images, preserving the detailed information of the images. L_{mod} minimizes the differences between different modalities to ensure that images of different modalities remain consistent in the feature space. L_{self} removes modality-specific information, enabling the model to focus on learning universal features of images rather than being influenced by specific modalities, thereby improving the model's generalization ability. For the specific details of each loss function, please refer to Reference[44]. The total loss L_{DN} is the weighted sum of the five loss functions:

$$L_{DN} = \lambda_{adv}\left(L_{adv}^{I} + L_{adv}^{I^m}\right) + \lambda_{\text{mod}}L_{\text{mod}} + \lambda_{rec}L_{rec} + \lambda_{self}L_{self} \tag{10}$$

where λ_{adv}, λ_{mod}, λ_{rec}, and λ_{self} are weighting parameters for different loss terms.

After training with the aforementioned loss function and images, the parameters of the encoder and decoder in the disentanglement network are effectively optimized, enabling the network to extract vessel information from CF and FFA images for use by the deformable network. It should be noted that to address the modality differences between CF and FFA images, this study adopted an independent training approach by training two separate disentanglement networks. Although the two networks share an identical architecture, their optimized parameters differ due to the distinct image modalities, allowing them to effectively adapt to the disentanglement requirements of both types of images.

- Deformation field generation
  The structure of deformable network is shown in Fig. 6, and its shape is very similar to that of Unet[51]. This paper adopts the registration field estimation network from[24] as the deformable network. The network takes the modality-free multimodal retinal images obtained from the disentangled network as input and outputs a deformation field. The network consists of five downsampling operations and three upsampling operations, resulting in a deformation field $F\prime$ with the shape $2 \times (w/4) \times (h/4)$, where h and w are the height and width of the image, respectively. Bilinear interpolation is then used to upsample the deformation field back to the original image size F.

The loss function of the deformable network consists of two parts: content loss and smoothness loss. The content loss uses the mean squared error between the fixed image and the warped moving image, which enables the content information of the disentangled modality-free retinal images to form good correspondences. Specifically, it is defined as:

$$L_{content} = MSE\left(STN\left(I_{mov}^{adn}, F\right), I_{fix}^{adn}\right) \tag{11}$$

where *MSE* is the mean squared error function, and *STN* [52] is the spatial transformation network that warps I_{mov}^{adn} according to the deformation field F. The smoothness loss calculates the absolute difference in displacement between adjacent pixels in the horizontal and vertical directions, and then takes the average of these differences over all pixels. This loss penalizes the discontinuity of the deformation field F and encourages

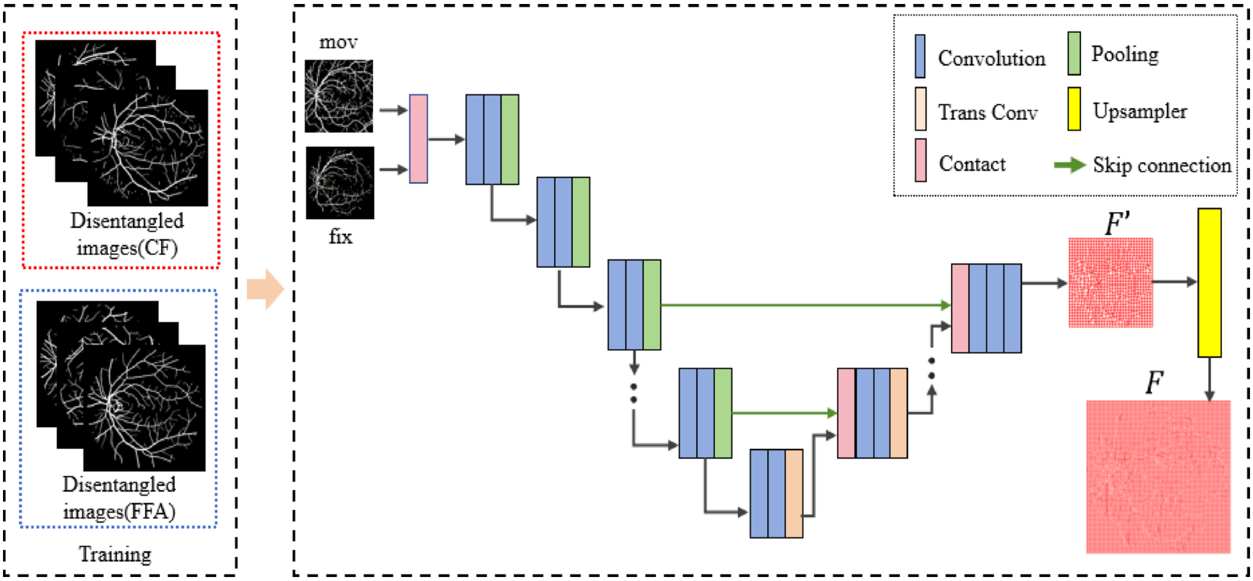


Fig. 6. The process of deformation field generation.

smaller displacement changes between adjacent pixels, making the deformation field smoother. The smoothness loss is defined as:

$$L_{smooth} = mean_{k,i,j}\left(|F_{k,i,j} - F_{k,i+1,j}| + |F_{k,i,j} - F_{k,i,j+1}|\right) \tag{12}$$

where k is the channel index of the deformation field, and i and j are the pixel indices of the image.

In addition to the aforementioned conventional smoothness constraint, we further introduce a second-order Laplacian regularization term into the loss function of the deformable network to further refine its output deformation field. The Laplacian operator is used to calculate the second-order derivative of an image or field and can capture higher-order changes.

The Laplacian penalty for the horizontal direction and the vertical direction is defined as:

$$\nabla_u^2 = -4u_{i,j} + u_{i,j+1} + u_{i,j-1} + u_{i+1,j} + u_{i-1,j} \tag{13}$$

$$\nabla_v^2 = -4v_{i,j} + v_{i,j+1} + v_{i,j-1} + v_{i+1,j} + v_{i-1,j} \tag{14}$$

where u and v represent the displacement in the horizontal and vertical directions, respectively. The Laplacian operator calculates the second-order derivative differences at each pixel location, penalizing larger changes.

The Laplacian penalty loss is:

$$L_{lap} = \frac{1}{H \times W}\sum_{i,j}\left(\nabla_u^2 + \nabla_v^2\right) \tag{15}$$

where H and W are the height and width of the image, respectively.

Therefore, the total loss of the deformable network in this study is:

$$L_{DF} = \lambda_{content}L_{content} + \lambda_{smooth}L_{smooth} + \lambda_{lap}L_{lap} \tag{16}$$

where $\lambda_{content}$, λ_{smooth}, and λ_{lap} are weighting parameters for different loss terms.

- Fine registration
  In the fine registration stage, the testing images from Dataset-1 images are processed through a trained disentangled network to extract vessel structures, which are subsequently subjected to an affine transformation obtained in the coarse registration stage and then fed into a trained deformable network to generate the final deformation fields. Fine registration is accomplished by applying the resulting deformation fields to the initially aligned images (i.e., the coarsely-registered images). This workflow is illustrated in Fig. 7.

The algorithm implementation details

This section introduces the implementation process of the proposed coarse-to-fine registration method in the form of pseudocode. The pseudocode of the method is shown in Table 1, and the specific steps are as follows:

Experiments and discussions

Experimental settings

Parameter settings

This section presents the experimental parameters for both registration stages. The coarse registration employed an SGD optimizer with 300 training epochs, using a batch size of 16, learning rate of 0.01, momentum of 0.937, and weight decay of 0.0005. For fine registration, the disentangled network was trained for 120 epochs with a batch size of 4, a learning rate of 1×10^{-4}, and a weight decay coefficient of 1×10^{-4}. The loss function employed weighting parameters λ_{adv}, λ_{rec}, λ_{mod}, and λ_{self}, set to 1, 20, 20, and 20, respectively. The deformable network was trained for 50 epochs with a batch size of 1 and a learning rate of 2×10^{-5}, using loss function weighting

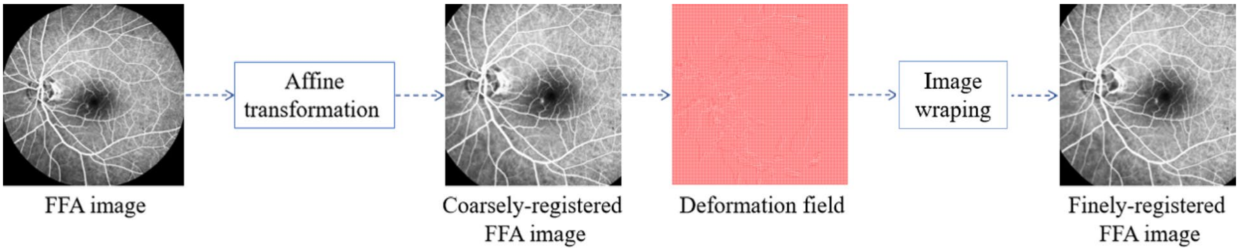


Fig. 7. The main process of fine registration: (1) Vessel structure extraction from the input images using a trained disentangled network; (2) Deformation field generation by processing the extracted structures through a trained deformable network; (3) Image warping by applying the computed deformation fields to achieve the finely-registered results.

```
Coarse Registration stage

① *Train the Yolov8-pose network*
    Input: $D^{C}_{Train}$, Epochs, SGD, Batch size, Learning rate, Momentum, Weight decay;
    for each epoch in Epochs do
        Shuffle( $D^{C}_{Train}$); // $D^{C}_{Train} \in$ Dataset-1
        for each batch in $D^{C}_{Train}$ do
            Predictions←Yolov8-pose( $I'_{CF\ \ FFA}$, $\theta_{Yolo}$); // ( $I'_{CF\ \ FFA}$, Annotation) $\in$ $D^{C}_{Train}$
            Calculate the $L_{Total\_Loss}$ using Predictions and Annotations;
            Update $\theta_{Yolo}$ via SGD;
        end
    end
    Output: $\theta^{*}_{Yolo}$;

② *Keypoint Optimization and Coarse Registration*
    Input: $D^{C}_{Test}$, $\theta^{*}_{Yolo}$; // $D^{C}_{Test} \in$ Dataset-1;
    $P^{k}_{CF}$, $P^{k}_{FFA}$←Yolov8-pose( $I''_{CF\_FFA}$, $\theta^{*}_{Yolo}$); // $I''_{CF\_FFA} \in D^{C}_{Test}$
    $P^{*}_{CF}$, $P^{*}_{FFA}$←KeypointOptimization( $P^{k}_{CF}$, $P^{k}_{FFA}$, $I''_{CF\_FFA}$);
    $T_{affine}$←AffineTransform( $P^{*}_{CF}$, $P^{*}_{FFA}$);
    $I_{FFA\ \ registered}$← $T_{affine}$( $I''_{FFA}$); // $I''_{FFA} \in D^{C}_{Test}$
    Output: $I_{FFA\_registered}$

Fine Registration Stage

① *Train the Disentangled Network*
    Input: $D^{F}_{Train}$, Epochs, Batch size, Learning rate, Weight decay, Adam, $\lambda_{adv}$, $\lambda_{rec}$, $\lambda_{mod}$, $\lambda_{self}$;
    for each epoch in Epochs do
        Shuffle( $D^{F}_{Train}$); // $D^{F}_{Train} \in$ Dataset-2
        for each batch in $D^{F}_{Train}$ do
            Predictions←Disentangled( $I_{CF}$, $I_{FFA}$, Pseudo labels); // ( $I_{CF}$, $I_{FFA}$, Pseudo labels) $\in$ $D^{F}_{Train}$
            Calculate the $L_{DN}$ using Formula(10);
            Update $\theta_{Disentangled}$ via Adam;
        end
    end
    Output: $\theta^{*}_{Disentangled}$;

② *Obtain the $D'_{vessel}$ Based on Dataset-1 and $\theta^{*}_{Disentangled}$*
    Input: $D^{C}_{Train}$, $D^{C}_{Test}$, $\theta^{*}_{Disentangled}$;
    $D_{vessel}$←Disentangled( $D^{C}_{Train}$, $D^{C}_{Test}$, $\theta^{*}_{Disentangled}$);
    $D'_{vessel}$← $T_{affine}$( $D_{vessel}$);
    Output: $D'_{vessel}$;

③ *Train the Deformable Network*
    Input: $D^{v}_{Train}$, Epochs, Batch size, Learning rate, Adam, $\lambda_{content}$, $\lambda_{smooth}$, $\lambda_{lap}$;
    for each epoch in Epochs do
        Shuffle( $D^{v}_{Train}$); // $D^{v}_{Train} \in D'_{vessel}$
        for each batch in $D^{v}_{Train}$ do
            Prediction←Deformable( $I^{v}_{CF}$, $I^{v}_{FFA}$, $\theta_{Deformable}$); // ( $I^{v}_{CF}$, $I^{v}_{FFA}$) $\in$ $D^{v}_{Train}$
            Calculate the $L_{DF}$ using Formula(11), (12), (15) and (16);
        end
    end
    Output: $\theta^{*}_{Deformable}$;

④ *Fine Registration*
    Input: $D^{v}_{Test}$, $D^{C}_{Test}$, $\theta^{*}_{Deformable}$;
    $F$←Deformable( $I^{v}_{CF}$, $I^{v}_{FFA\_registered}$, $\theta^{*}_{Deformable}$); // ( $I^{v}_{CF}$, $I^{v}_{FFA\_registered}$) $\in$ $D^{v}_{Test}$, $D^{v}_{Test} \in D'_{vessel}$
    $I_{FFA\_fine}$←wrap($F$, $I_{FFA\_registered}$); // $I_{FFA\_registered} \in D^{C}_{Test}$
    Output: $I_{FFA\_fine}$.
```

Table 1. The Pseudocode for the proposed coarse-to-fine registration method.

parameters $\lambda_{content}$, λ_{smooth}, and λ_{lap} set to 2×10^{-3}, 2×10^{-5}, and 1×10^{-5}, respectively. Both networks were implemented in the PyTorch framework and optimized using the Adam optimizer.

Evaluation metric

In this study, the performance of the proposed registration method is evaluated using Dice coefficient, Dice_s coefficient, and the percentage of non-positive values in the determinant of the Jacobian matrix of deformation fields. Dice coefficient and Dice_s coefficient are used to indicate the accuracy of registration, while the percentage of non-positive values in the determinant of the Jacobian matrix of deformation fields is used to characterize the smoothness of the deformation field.

The Dice coefficient measures the overlap between two sets and has increasingly been used in recent years to evaluate the accuracy of image registration, indicating the overlap between the target and reference regions after registration. The formula for calculating the Dice coefficient is as follows:

$$Dice\left(S_1, S_2\right) = \frac{2\times\sum\left(S_1 \odot S_2\right)}{\sum S_1 + \sum S_2} \tag{17}$$

where S_1 and S_2 are the binary segmentation results of the two images, and $\odot$ denotes the element-wise product. The Dice coefficient ranges from 0 to 1, with higher values indicating greater overlap.

The Dice_s coefficient is a differentiable version of the Dice coefficient, suitable for evaluating probabilistic segmentation maps. The formula for calculating the soft Dice coefficient is as follows:

$$Dice_s\left(P_1, P_2\right) = \frac{2\times\sum ele\_\min\left(P_1, P_2\right)}{\sum P_1 + \sum P_2} \tag{18}$$

where P_1 and P_2 are the probabilistic segmentation maps of the two images, and $ele\_\min\left(P_1, P_2\right)$ represents the element-wise minimum. The soft Dice coefficient also ranges from 0 to 1, with higher values indicating greater overlap.

Percentage of non-positive values in determinant of Jacobian matrix of deformation fields ($\%of\ |J_F| \leq 0$) is added to assess the smoothing effect of the deformation field after introducing the Laplacian penalty loss, which is defined as:

$$\%of\ |J_F| \leq 0 = \frac{1}{|\Omega|}\sum_{\Omega} 1\left(\det\left(F\right) \leq 0\right) \tag{19}$$

where F is the deformation field, $\det\left(F\right)$ is the determinant of the Jacobian matrix of the deformation field, and Ω is the entire image domain. $1\left(\det\left(F\right) \leq 0\right)$ is an indicator function that takes the value 1 when $\det\left(F\right) \leq 0$ and 0 otherwise. $\sum_{\Omega} 1\left(\det\left(F\right) \leq 0\right)$ calculates the number of pixels in the entire image domain where $\det\left(F\right) \leq 0$. $\det\left(F\right)$ provides information about the deformation at each pixel location: when $\det\left(F\right) > 1$, it indicates expansion at that pixel location; when $\det\left(F\right) = 1$, it indicates no change; when $0 < \det\left(F\right) < 1$, it indicates contraction; and when $\det\left(F\right) \leq 0$, it indicates folding, which violates diffeomorphic properties.

Discussion

This section will evaluate the performance of the proposed multimodal retinal image registration method for CSCR, covering both coarse and fine registration stages.

Discussion on the coarse registration stage

1. Qualitative Analysis

To evaluate the effectiveness of our proposed three-step coarse registration strategy, we conducted comparative experiments with two representative registration methods (B-COSFIRE[53] + SIFT[54] and SURF-PIIFD-RPM[8]). The key differences between these methods lie primarily in their feature extraction algorithms, matching strategies, and mismatch elimination techniques.

- B-COSFIRE[53] + SIFT[54]: This method integrates multiple techniques for multimodal retinal image registration. Initially, B-COSFIRE filters preprocess the retinal vessels, followed by refinement through threshold segmentation and skeletonization. Subsequently, the SIFT algorithm detects salient feature points, while brute-force matching establishes correspondences. The RANSAC algorithm then automatically removes mismatches, with final alignment achieved through affine transformation.
- SURF-PIIFD-RPM[8]: This method employs SURF for local feature point detection, extracts feature descriptors using PIIFD (Partial Intensity Invariant Feature Descriptor), eliminates mismatches via the RPM (Robust Point Matching) algorithm, and ultimately computes affine transformation parameters through weighted least squares to achieve final registration.

Figure 8 visually compares the registration results of different methods for CSCR multimodal retinal images. The first row shows the CF and FFA images to be registered, as well as the checkerboard overlay and local magnified views in their unregistered state. The second and third rows show the registration effects of the methods B-COSFIRE + SIFT and SURF-PIIFD-RPM, including the detected feature points and the registered checkerboard images. From the results in the first column of Fig. 8, B-COSFIRE + SIFT detects more keypoints in CF and FFA images. However, it fails to establish matches for many point pairs, and some of the established matches are incorrect. These issues place B-COSFIRE + SIFT at a certain disadvantage compared to both the SURF-PIIFD-RPM and the proposed three-step coarse registration strategy. Although the SURF-PIIFD-RPM method demonstrates better outlier rejection than B-COSFIRE + SIFT, its matched keypoints are excessively clustered. This prevents the computation of a suitable affine transformation matrix, resulting in only partial vessel alignment and ultimately leading to registration failure. The fourth row presents the registration results of our proposed three-step coarse registration strategy, with its first column showing the keypoint pairs predicted by YOLOv8 and the three pairs selected for the affine matrix computation. Visibly, the keypoints are accurately predicted, well-distributed, and all are correctly matched, which provides a reliable reference for the subsequent registration. Besides, a comparison of the locally magnified regions marked by red and blue boxes in the third column of Fig. 8 clearly shows that both the B-COSFIRE + SIFT and SURF-PIIFD-RPM methods perform poorly in local registration. In contrast, the proposed strategy achieves significantly better registration performance. In conclusion, the visualization registration results clearly indicate that our method successfully achieves comprehensive global alignment, with retinal vessels exhibiting precise spatial correspondence.

2. Quantitative Analysis

Figure 8 intuitively demonstrates the effectiveness of the proposed three-step coarse registration strategy in the coarse registration stage of multimodal retinal images for CSCR. To further verify its advantages, we conduct a quantitative performance analysis. As shown in Table 2, using the pre-registration Dice and Dice_s scores as baselines, the B-COSFIRE + SIFT and SURF-PIIFD-RPM methods achieve scores of 0.3908 and 0.2976, and 0.4301 and 0.3509, respectively. These represent improvements of 0.2239 and 0.0614, and 0.2632 and 0.1147 over the baseline values, demonstrating that both methods possess a partial ability to align multimodal fundus images. Meanwhile, compared with B-COSFIRE + SIFT, SURF-PIIFD-RPM demonstrates superior performance, as reflected in its leads of 0.0393 and 0.0533 in the two aforementioned metrics. Notably, our proposed CRS strategy yields Dice and Dice_s scores of 0.5023 and 0.3940, respectively, representing improve-

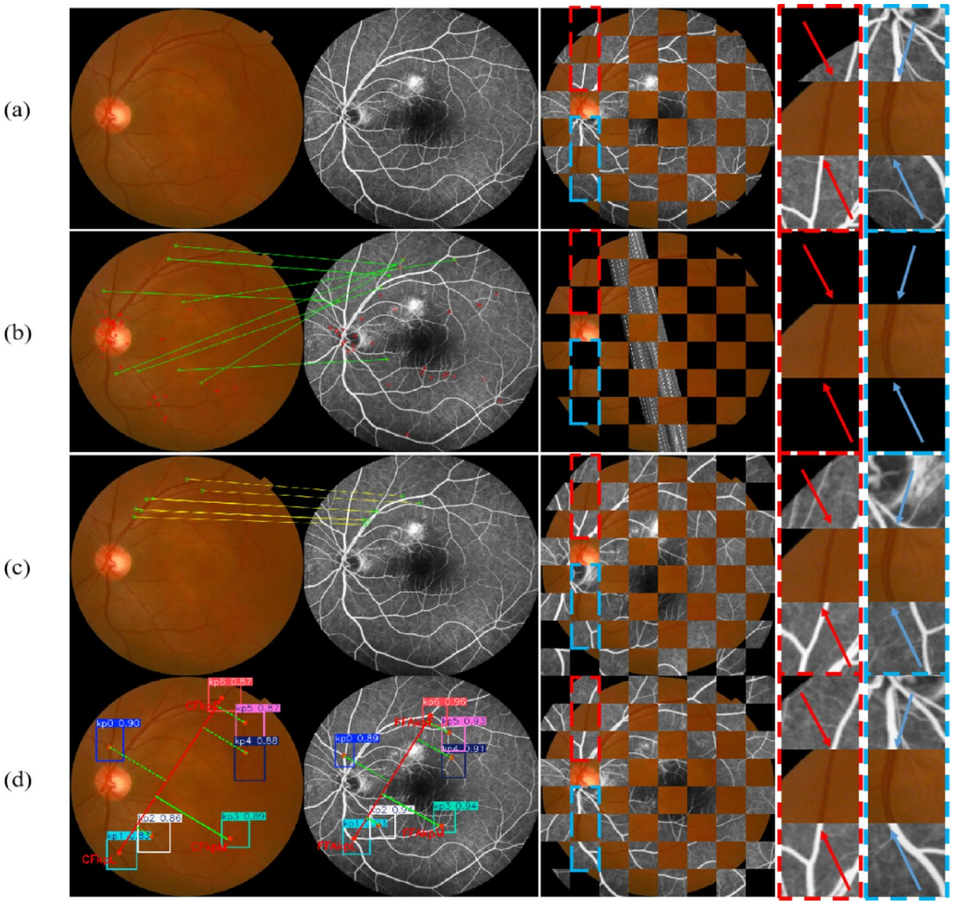


Fig. 8. Qualitative results of different registration methods: (**a**) Before registration, (**b**) B-COSFIRE + SIFT, (**c**) SURF-PIIFD-RPM, and (**d**) The proposed three-step coarse registration strategy. *Note*: The three columns show the images to be registered, checkerboard overlays, and local magnified views, with row 1 showing the pre-registration state and rows 2–4 showing the post-registration results.

Methods	Dice↑(± std)	Dice_s↑(± std)
Before registration	0.1669 (± 0.0238)	0.2362 (± 0.0096)
B-COSFIRE[53] + SIFT[54]	0.3908 (± 0.2571)	0.2976 (± 0.1847)
SURF-PIIFD-RPM[8]	0.4301 (± 0.1931)	0.3509 (± 0.0978)
CRS	0.5023 (± 0.1262)	0.3940 (± 0.0571)

Table 2. Quantitative results of different registration methods. CRS is the abbreviation for the proposed three-step coarse registration strategy.

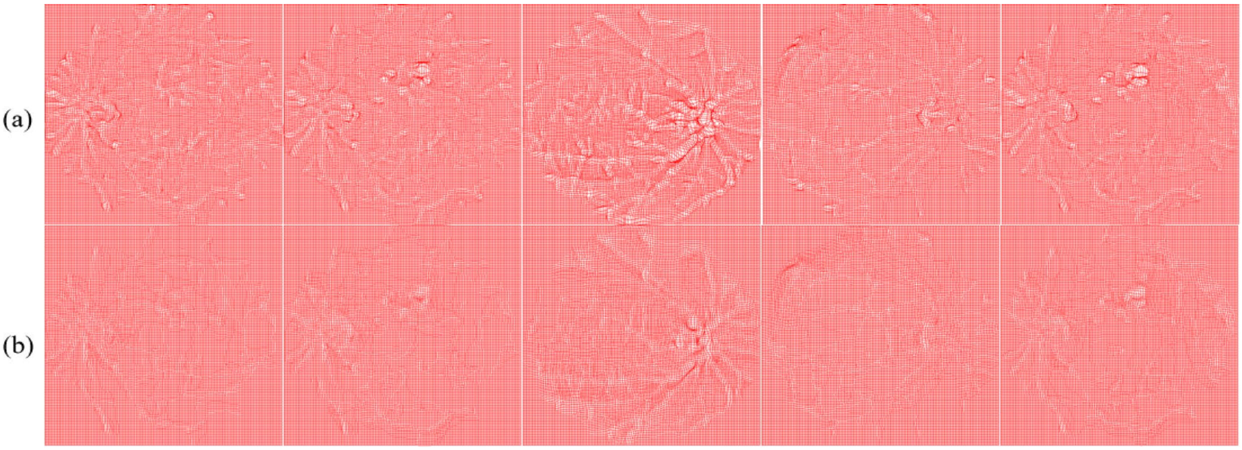


Fig. 9. Visual comparison between the deformation fields: (**a**) L_{DF} without the Laplacian penalty loss, and (**b**) L_{DF} with the Laplacian penalty loss. *Note*: the figure only shows the deformation fields of four samples, with each column representing one sample.

ments of 0.3354 and 0.1578 over the baseline values, thus confirming its effectiveness. Furthermore, compared with typical rigid registration methods, CRS shows significantly better performance, outperforming the SURF-PIIFD-RPM and B-COSFIRE + SIFT methods by 0.0722 and 0.0431, and 0.1115 and 0.0964 in the two metrics respectively. These observations are consistent with the visual registration results in Fig. 8, further confirming the superiority of our proposed strategy for CSCR multimodal retinal image registration.

Discussion on the fine registration stage

1. Ablation Study

This section presents ablation studies to evaluate the necessity of both the Laplacian penalty loss and the coarse registration stage, and then provides comprehensive qualitative and quantitative validation of the proposed dual-component fine registration strategy for CSCR multimodal retinal images.

- The necessity of Laplacian penalty loss
  Figure 9 presents a visual comparison between the deformation fields before and after applying the Laplacian penalty loss, using four samples as examples. The first row displays the deformation fields without the Laplacian penalty loss, while the second row shows the results after its application. It can be visually observed that when the Laplacian penalty is not applied, the deformation fields in the first row exhibit significant folding issues. However, after introducing the penalty, the second row results show a notable reduction in folding regions, with the deformation fields demonstrating improved continuity and smoothness overall. This demonstrates that the Laplacian penalty loss effectively suppresses unnecessary local distortions, improves the quality of the deformation field, and consequently enhances the accuracy and reliability of the registration.

In addition to the intuitive visualization of the deformation fields shown in Fig. 9, we also computed the percentage of non-positive Jacobian determinant values across the deformation field for each testing image pair, aiming to quantitatively evaluate the effect of the Laplacian penalty loss. As shown in Fig. 10, the introduction of Laplacian penalty loss leads to a significant reduction in the percentage of non-positive values in the Jacobian determinant of deformation fields across all 69 testing samples. The majority of samples reach 0% for this metric, while the remaining cases show values below 0.25%, representing a substantial improvement compared to the results without Laplacian penalty loss. These findings clearly demonstrate the effectiveness of introducing Laplacian penalty loss in this work.

- The necessity of the coarse registration stage
  Figure 11 shows the registration effects under four different configurations: (a) fine registration directly without coarse registration and without Laplacian penalty loss; (b) adding Laplacian penalty loss to configuration (a); (c) fine registration after coarse registration but without Laplacian penalty loss; (d) adding Laplacian penalty loss to configuration (c). The images shown from left to right are the deformation field, the transformed FFA image, the checkerboard image, and the locally magnified image. Quantitative results indicate that registration performance is noticeably inferior without the coarse registration stage, while substantial improvements are observed when it is applied. Additionally, we explored the joint effect of the Laplacian penalty loss and coarse registration, which significantly improves vessel spatial alignment in multimodal retinal images.

In terms of quantitative evaluation, the data in Table 3 further corroborate the above conclusions. The experimental group without coarse registration and without the introduction of Laplacian penalty loss performs the worst, with Dice and Dice$_s$ coefficients of only 0.4580 and 0.3792, respectively, and a high $\%of\ |J_F|$ of 1.2345, indicating severe deformation field distortion. It is worth noting that although introducing Laplacian penalty loss alone in the fine registration stage can significantly reduce $\%of\ |J_F|$ to 0.0186, the Dice and Dice$_s$ coefficients actually decrease, still failing to achieve the desired registration outcome. In contrast, the strategy of using coarse registration followed by fine registration significantly improves registration accuracy, with Dice and Dice$_s$ coefficients far superior to those without coarse registration. On this basis, the introduction of Laplacian penalty loss resulting in a slight decrease in Dice coefficient by 0.0248, but an increase in Dice$_s$ coefficient by 0.0042, while $\%of\ |J_F|$ is reduced by 0.3007, effectively reducing deformation field distortion and ultimately achieving more accurate registration results.

These experimental results fully demonstrate the significant value of coarse registration as a preprocessing stage from both qualitative and quantitative perspectives. It not only provides a good initial alignment for subsequent fine registration but also effectively constrains the rationality of the deformation field, thereby ensuring the accuracy and reliability of the final registration results.

2. Qualitative Analysis

Based on the above discussion, this section compares the proposed coarse-to-fine registration method with other multimodal retinal image deformable registration approaches from both qualitative and quantitative perspectives. The two comparison methods are Phase[55] + MIND[56], a non-deep learning-based approach, and RetinaSegReg[24], a deep learning-based method that utilizes style transfer.

- Phase[55] + MIND[56]: The method begins by extracting phase maps from multimodal images, followed by the computation of MIND based on the extracted phase information. Specifically, the Fourier transform is applied to capture phase features, which are known to be highly sensitive to structural variations within the images. The phase differences within local neighborhoods are then encoded to construct MIND descriptors that effectively characterize local structural properties. Finally, non-rigid registration is accomplished by minimizing the difference between the MIND descriptors of the image pair.
- RetinaSegReg [24]: This method represents an innovative deep learning-based approach to multimodal retinal image registration. Inspired by the concept of style transfer, it aims to improve vessel structure alignment by simulating style transformations between different modalities. The method first employs a style transfer network to segment blood vessels in multimodal retinal images. These segmentation results are

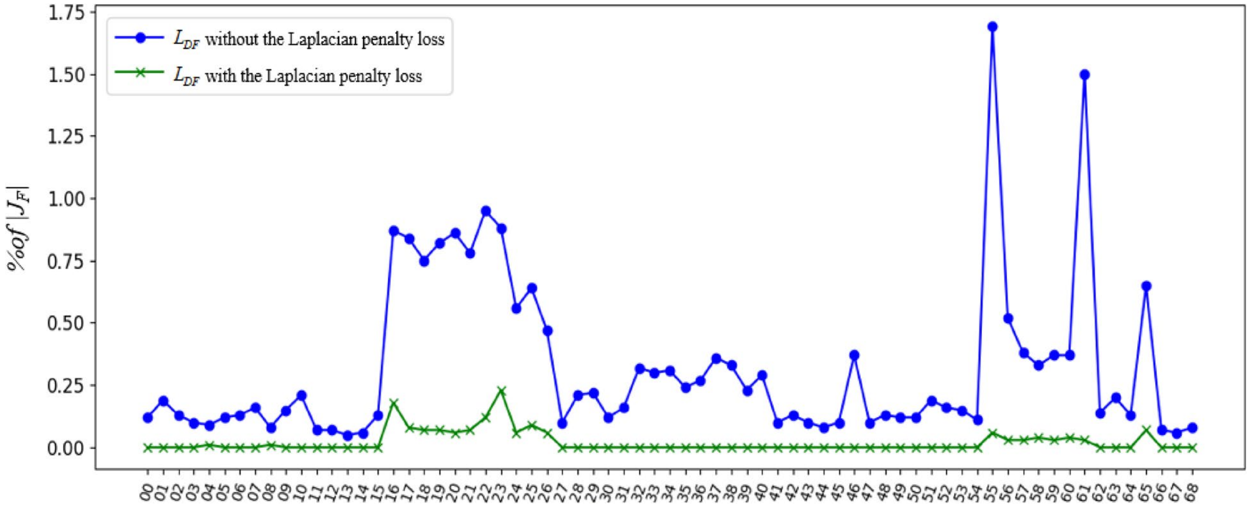


Fig. 10. Quantitative comparison based on the $\%of\ |J_F|$

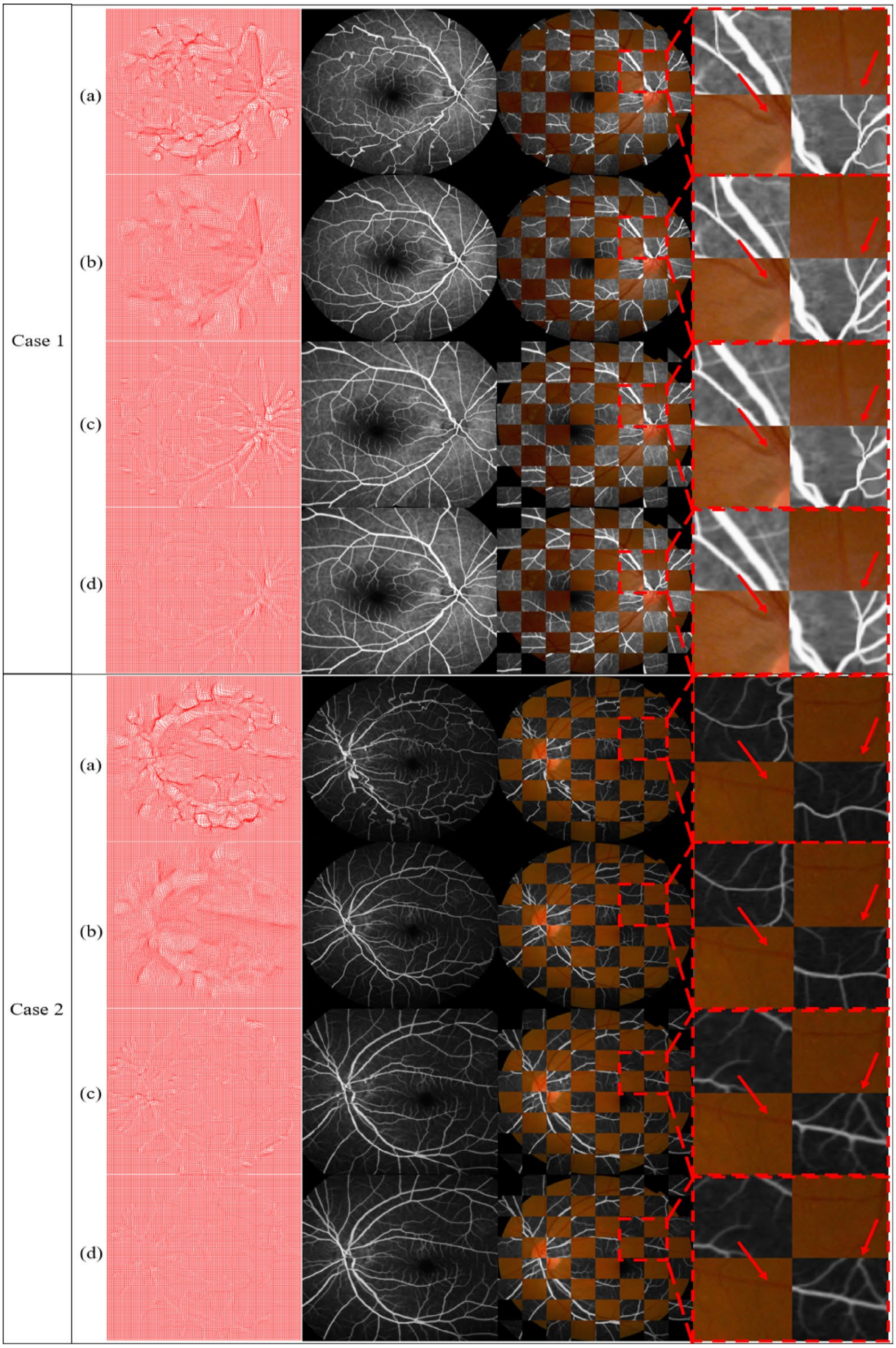


Fig. 11. The registration effects under four different configurations: (**a**) fine registration directly without coarse registration and without Laplacian penalty loss, (**b**) adding Laplacian penalty loss to configuration (a), (**c**) fine registration after coarse registration but without Laplacian penalty loss, and (**d**) adding Laplacian penalty loss to configuration (c). *Note*: The images shown from left to right are the deformation fields, the transformed FFA images, the checkerboard images, and the locally magnified views.

then used as guidance for precise image registration. The strength of this approach lies in its ability to effectively reduce inherent cross-modality differences, thereby enhancing registration accuracy.

Configurations	Dice↑(±std)	Dice$_s$↑(±std)	$\%of\ \|J_F\|$↓(±std)
w/o C and L	0.4580(±0.0478)	0.3792(±0.0161)	1.2345(±0.1097)
w/o C	0.3592(±0.0519)	0.3257(±0.0200)	0.0186(±0.0007)
w/o L	0.7007(±0.0924)	0.4935(±0.0309)	0.3216(±0.3291)
w/ L	0.6759(±0.1089)	0.4977(±0.0393)	0.0209(±0.0426)

Table 3. The quantitative assessment of the impact of coarse registration on fine registration performance. The abbreviations "w/" and "w/o" represent "with" and "without", respectively, while "C" and "L" denote "coarse registration" and "Laplacian penalty loss", respectively.

Figure 12 illustrates the visual registration results of different methods based on two representative examples. Figure 12a shows the result of the coarse registration stage proposed in this study, where the images include the CF image, the registered FFA image, the checkerboard image, and a locally magnified view. Figures 12b, c, and d present the results of Phase + MIND, RetinaSegReg, and our proposed coarse-to-fine registration method, respectively, each showing the deformation field, the registered FFA image, the checkerboard image, and a magnified view.

From the checkerboard image in Fig. 12a, it can be clearly seen that, the multimodal retinal images are well aligned in terms of blood vessels after coarse registration. However, as indicated by the red arrows in the local magnified image, there are still minor deviations. This indicates that while rigid registration can handle global deformation, it remains insufficient for addressing local registration issues, necessitating non-rigid registration to resolve local deformation problems. In addition, from the perspective of deformation field smoothness, the deformation field in Fig. 12b is the smoothest, and the non-rigidly transformed FFA image also exhibits the most natural gradual transition. This result demonstrates the advantage of traditional method in maintaining deformation field smoothness, although the issue of low computational efficiency of this method remains non-negligible. In Fig. 12c, the edges of the FFA image are noticeably folded inward, indicating that the style transfer algorithm has introduced unreasonable distortions in local areas of the deformation field, compromising the anatomical plausibility of the registration result. In Fig. 12d, the local magnified image, as indicated by the red arrow, shows well-aligned retinal vessels, demonstrating the effectiveness of the proposed method in local fine registration. It is particularly noteworthy that while the deformation field generated by RetinalSegReg exhibits some folding, the deformation field produced by our method is comparatively smoother. This smoothness avoids unnecessary local distortions, thereby ensuring the anatomical credibility of the registration results. Overall, the proposed coarse-to-fine registration method shows certain advantages and competitive performance, both when compared to the three-step coarse registration strategy and to the Phase + MIND and RetinalSegReg methods.

3. Quantitative Analysis

Table 2 and Fig. 8 validate the effectiveness of the proposed coarse registration strategy. Building on these findings and the visual registration results shown in Fig. 12, this section further quantitatively investigates the impact of introducing a dual-component fine registration strategy on the registration results. As shown in Table 4, the proposed coarse-to-fine (C2F) method achieves final Dice and Dices coefficients of 0.6759 and 0.4977, representing improvements of 0.1736 and 0.1037 over the previous coarse registration strategy (i.e., CRS), respectively, thereby further confirming the effectiveness and necessity of the fine registration strategy. Compared to the Phase + MIND, the proposed coarse-to-fine (C2F) method shows comparable performance in the Dice$_s$ coefficient but has a higher Dice coefficient by 0.0244. In comparison with the RetinalRegSeg method, although C2F has a slightly lower Dice$_s$ coefficient by 0.01023, it has a higher Dice coefficient by 0.0274 and a lower $\%of\ |J_F|$ by 0.1258, indicating less distortion in the images registered by the C2F method. The comprehensive comparison indicates that the proposed method achieves superior registration accuracy. Overall, the quantitative results combined with the visual registration results in Fig. 12 demonstrate that the proposed coarse-to-fine registration method achieves promising registration performance and possesses significant competitive potential.

Limitations and future work

The experimental analysis and discussion of the aforementioned coarse and fine registration demonstrate that the proposed method achieves promising registration results for clinical multimodal fundus images of CSCR. This provides a potential solution for laser preoperative auxiliary registration in CSCR treatment. While substantial progress has been made in current research, several challenges require further attention. First, the present fine registration approach adopts a non-end-to-end architecture, maintaining relative independence between the disentanglement and registration processes. Future work should focus on developing an end-to-end unified framework that integrates both components. Second, when processing CF images, suboptimal imaging quality of certain major vessels results in punctate or linear discontinuities within the vessels following disentanglement, severely compromising registration accuracy. Consequently, developing CF-specific vessel enhancement algorithms emerges as a crucial research priority. These advancements promise to substantially improve both the robustness and clinical utility of the CSCR multimodal retinal image registration method.

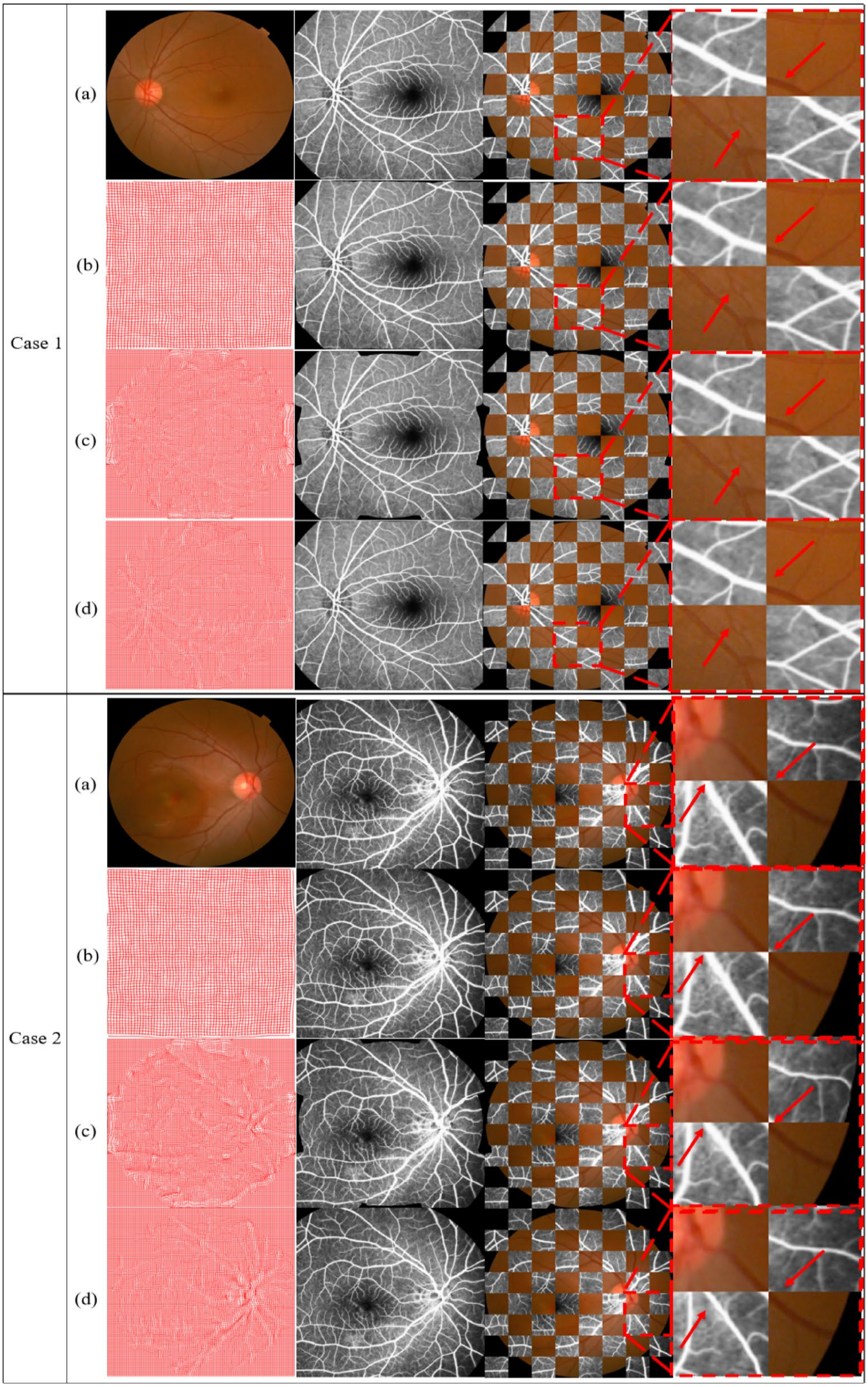


Fig. 12. Qualitative results of different registration methods: (**a**) The proposed three-step coarse registration strategy, (**b**) Phase + MIND, (c) RetinalSegReg, and (d) The proposed coarse-to-fine registration. *Note*: The first column shows the original CF images or deformation fields, with columns 2 to 4 displaying the registered images, checkerboard images, and local magnified views, respectively.

Methods	Dice↑(± std)	Dice_s↑(± std)	$\%of\ \lvert J_F\rvert$↓(± std)
CRS	0.5023(± 0.1262)	0.3940(± 0.0571)	/
Phase[55] + MIND[56]	0.6515(± 0.1283)	0.5008(± 0.0485)	0.0000(± 0.0000)
RetinalSegReg[24]	0.6485(± 0.0947)	0.5100(± 0.0305)	0.1467(± 0.4456)
C2F	0.6759(± 0.1089)	0.4977(± 0.0393)	0.0209(± 0.0426)

Table 4. The quantitative comparison of different methods. CRS and C2F are the abbreviations for the proposed three-step coarse registration strategy and the proposed coarse-to-fine registration method, respectively.

Conclusions

To address the challenges in preoperative CSCR registration arising from ophthalmologists' manual operations and the limitations of existing rigid and non-rigid registration methods, this study proposes a coarse-to-fine registration method for multimodal retinal images. The methodology comprises two key parts: (1) a three-step coarse registration strategy employing the YOLOv8-pose network to unify keypoint detection and matching, with keypoints further optimized through a post-processing technique, followed by affine transformation for preliminary multimodal image alignment; and (2) a dual-component fine registration strategy that implements a disentanglement learning approach to preserve vessel structures while eliminating modality-specific discrepancies, ultimately achieving refinement of the coarsely-registered images through a deformable network. Both extensive qualitative and quantitative experiments have validated the effectiveness of our proposed method, demonstrating its potential to provide technical support for multimodal retinal image registration in CSCR preoperative procedures. Future research will be conducted to further enhance the performance of this methodology.

Data availability

The datasets used during the current study are available from the corresponding authors upon reasonable request.

Received: 6 July 2025; Accepted: 27 November 2025
Published online: 08 December 2025

References

1. Wu, Z. et al. Clinical characteristics, treatment, and outcomes of nivolumab-induced uveitis. *Immunopharmacol. Immunotoxicol.* **47**(2), 222–227 (2025).
2. Wu, Z., Sun, W. & Wang, C. Clinical characteristics, treatment, and outcomes of pembrolizumab-induced uveitis. *Invest. New Drugs* **42**(5), 510–517 (2024).
3. Legg, P. A. et al. Improving accuracy and efficiency of mutual information for multi-modal retinal image registration using adaptive probability density estimation. *Comput. Med. Imaging Graph.* **37**(7–8), 597–606 (2013).
4. Reel P. S., Dooley L. S., Wong K. C. P. et al. Robust retinal image registration using expectation maximisation with mutual information. 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP) IEEE, 1118–1121 (2013).
5. Lange A., Heldmann S. Multilevel 2D-3D intensity-based image registration. Biomedical Image Registration: 9th International Workshop, WBIR 2020, Portorož, Slovenia, December 1–2, 2020, Proceedings 9. Springer International Publishing, 57–66 (2020).
6. Yang, G. et al. Registration of challenging image pairs: Initialization, estimation, and decision. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(11), 1973–1989 (2007).
7. Ghassabi, Z. et al. An efficient approach for robust multimodal retinal image registration based on UR-SIFT features and PIIFD descriptors. *EURASIP J. Image Video Process.* **2013**(1), 1–16 (2013).
8. Wang, G. et al. Robust point matching method for multimodal retinal image registration. *Biomed. Signal Process. Control* **19**, 68–76 (2015).
9. Chen L., Xiang Y., Chen Y. J. et al. Retinal image registration using bifurcation structures. IEEE International Conference on Image Processing, 2169–2172(2011).
10. Wang Y., Zhang J., An C. et al. A segmentation based robust deep learning framework for multimodal retinal image registration. ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1369–1373 (2020).
11. Wang, Y. et al. Robust content-adaptive global registration for multimodal retinal images using weakly supervised deep-learning framework. *IEEE Trans. Image Process.* **30**, 3167–3178 (2021).
12. Ali Z., Hussain T., Su C. et al. A Novel Hybrid Signal Processing Based Deep Learning Method for Cyber-Physical Resilient Harbor Integrated Shipboard Microgrids. IEEE Transactions on Industry Applications. 1–19 (2025).
13. Bakkouri, I. et al. BG-3DM2F: Bidirectional gated 3D multi-scale feature fusion for Alzheimer's disease diagnosis. *Multimedia Tools and Applications.* **81**, 10743–10776 (2022).
14. Hussain, T. et al. EFFResNet-ViT: A fusion-based convolutional and vision transformer model for explainable medical image classification. *IEEE Access.* **13**, 54040–54068 (2025).
15. Zhang, Z. et al. Deep learning and radiomics-based approach to meningioma grading: exploring the potential value of peritumoral edema regions. *Phys. Med. Biol.* **69**, 105002 (2024).
16. Zhu, C. et al. AMSFuse: Adaptive Multi-Scale Feature Fusion Network for Diabetic Retinopathy Classification. *Comput Mater. Continua.* **82**, 5153–5167 (2025).
17. Bakkouri, I. & Afdel, K. MLCA2F: Multi-level context attentional feature fusion for COVID-19 lesion segmentation from CT scans. *SIViP* **17**, 1181–1188 (2023).
18. Hussain, T. et al. DCSSGA-UNet: Biomedical image segmentation with DenseNet channel spatial and Semantic Guidance Attention. *Knowl.-Based Syst.* **314**, 113233 (2025).
19. Ji, Z. et al. BGRD-TransUNet: A Novel TransUNet-Based Model for Ultrasound Breast Lesion Segmentation. *IEEE Access.* **12**, 31182–31196 (2024).

20. Yu J., Qin J., Xiang J. et al. Trans-UNeter: A new Decoder of TransUNet for Medical Image Segmentation. 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2338–2341 (2023).
21. Zhang, Z. et al. A novel deep learning model for medical image segmentation with convolutional neural network and transformer. *Interdisciplinary Sci. Comput. Life Sci.* **15**, 663–677 (2023).
22. Umirzakova, S. et al. Enhancing the super-resolution of medical images: Introducing the deep residual feature distillation channel attention network for optimized performance and efficiency. *Bioengineering* **10**(11), 1332 (2023).
23. Lee J. A., Liu P., Cheng J. et al. A deep step pattern representation for multimodal retinal image registration. Proceedings of the IEEE/CVF International Conference on Computer Vision, 5076–5085 (2019).
24. Zhang J., An C., Dai J. et al. Joint vessel segmentation and deformable registration on multi-modal retinal images based on style transfer. 2019 IEEE International conference on image processing (ICIP), 839–843 (2019).
25. Zhang J., Wang Y., Dai J., et al. . IEEE Transactions on Image Processing. 31, 823-838(2022).
26. Martínez-Río, J. et al. Deformable registration of multimodal retinal images using a weakly supervised deep learning approach. *Neural Comput. Appl.* **35**(20), 14779–14797 (2023).
27. Santarossa, M. et al. MedRegNet: Unsupervised multimodal retinal image registration with GANs and ranking loss. Medical Imaging 2022: Image Processing. *SPIE* **12032**, 321–333 (2022).
28. Chih-Hung W., Tzuching L., Guanru C. et al. Deep Learning-based Diagnosis and Localization of Pneumothorax on Portable Supine Chest X-ray in Intensive and Emergency Medicine: A Retrospective Study. Journal of Medical Systems. (2023).
29. Bakkouri, I. & Afdel, K. Computer-aided diagnosis (CAD) system based on multi-layer feature fusion network for skin lesion recognition in dermoscopy images. *Multimed. Tools Appl.* **79**, 20483–20518 (2020).
30. Bakkouri, I. & Afdel, K. Multi-scale CNN based on region proposals for efficient breast abnormality recognition. *Multimed. Tools Appl.* **78**, 12939–12960 (2019).
31. Ye, Y. et al. A hybrid bioelectronic retina-probe interface for object recognition. *Biosens. Bioelectron.* **279**, 117408 (2025).
32. Ali Z., Hussain T., Su C. et al. Deep Learning-Driven Cyber Attack Detection Framework in DC Shipboard Microgrids System for Enhancing Maritime Transportation Security. IEEE Transactions on Intelligent Transportation Systems.1–12 (2025).
33. Ali Z., Hussain T., Su C. et al. A Novel Intelligent Intrusion Detection and Prevention Framework for Shore-Ship Hybrid AC/DC Microgrids Under Power Quality Disturbances. 2025 IEEE Industry Applications Society Annual Meeting (IAS), 1–7 (2025).
34. Jian, S. U., Fang, W. & Wei, Z. An improved YOLOv7-tiny algorithm for vehicle and pedestrian detection with occlusion in autonomous driving. *Chin. J. Electron.* **34**(1), 1–13 (2025).
35. Qian, L., Zheng, Y. & Liu, Z. X. Lightweight ship target detection algorithm based on improved YOLOv5s. *J. Real-Time Image Process.* **21**(1), 3.1-3.15 (2024).
36. He, L. et al. Research on object detection and recognition in remote sensing images based on YOLOv11. *Sci. Rep.* **15**, 14032 (2025).
37. Hou, Y. et al. A rapid detection method for wheat seedling leaf number in complex field scenarios based on improved YOLOv8. *Smart Agriculture.* **6**(4), 128–137 (2024).
38. Al-antari, M. A. et al. A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification. *Int. J. Med. Inform.* **117**, 44–54 (2018).
39. Huang X., Liu M. Y., Belongie S. et al. Multimodal Unsupervised Image-to-Image Translation. Proceedings of the European conference on computer vision (ECCV), 172–189 (2018).
40. Lee H. Y., Tseng H. Y., Huang J. B. et al. Diverse image-to-image translation via disentangled representations. Proceedings of the European conference on computer vision (ECCV), 35–51 (2018).
41. Huang P., Hu S., Peng B. et al. Robustly Optimized Deep Feature Decoupling Network for Fatty Liver Diseases Detection. arXiv, 2024.
42. Jin Z., Wang C., Luo X. Colorization-Inspired Customized Low-Light Image Enhancement by a Decoupled Network. IEEE Transactions on Neural Networks and Learning Systems.1–14 (2024).
43. Qin C., Shi B., Liao R. et al. Unsupervised deformable registration for multimodal images via disentangled representations. International Conference on Information Processing in Medical Imaging. Cham: Springer International Publishing, 249–261 (2019).
44. Liao, H. et al. ADN: Artifact disentanglement network for unsupervised metal artifact reduction. *IEEE Trans. Med. Imaging* **39**(3), 634–643 (2019).
45. He, W. et al. Implementation and validation of a three-dimensional deformable registration algorithm for targeted prostate cancer radiotherapy - ScienceDirect. *Int. J. Radiat. Oncol. Biol. Phys.* **61**(3), 725–735 (2005).
46. Mohamed, A. et al. Deformable registration of brain tumor images via a statistical model of tumor-induced deformation. *Med. Image Anal.* **10**(5), 752–763 (2006).
47. Michael, V. B., Joanne, L. M. & Cynthia, L. E. Effect of breathing motion on radiotherapy dose accumulation in the abdomen using deformable registration. *Int. J. Radiat. Oncol. Biol. Phys.* **80**(1), 265–272 (2011).
48. Perez-Rovira A., Trucco E., Wilson P. et al. Deformable registration of retinal fluorescein angiogram sequences using vasculature structures. Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference, 2010, 4383–4386 (2010).
49. Staal, J. et al. Ridge-based vessel segmentation in color images of the retina. *IEEE Trans. Med. Imaging* **23**(4), 501–509 (2004).
50. Jocher G., Chaurasia A., Qiu J. Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics. (2023)
51. Ronneberger O., Fischer P., Brox T. U-net: Convolutional networks for biomedical image segmentation. Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18. Springer International Publishing, 2015, 234–241 (2015).
52. Jaderberg M., Simonyan K., Zisserman A. et al. Spatial Transformer Networks. MIT Press, (2015).
53. Azzopardi, G. et al. Trainable COSFIRE filters for vessel delineation with application to retinal images. *Med. Image Anal.* **19**(1), 46–57 (2015).
54. Lowe D. G. Object recognition from local scale-invariant features. In Proceedings of the seventh IEEE international conference on computer vision. IEEE, 2, 1150–1157(1999).
55. Felsberg, M. & Sommer, G. The monogenic signal. *IEEE Trans. Signal Process.* **49**(12), 3136–3144 (2001).
56. Heinrich, M. P. et al. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Med. Image Anal.* **16**(7), 1423–2143 (2012).

Acknowledgements

We sincerely thank the cooperation unit for providing the CSCR images for this scientific research. We also greatly appreciate the editors and reviewers for their time and attention to our paper.

Author contributions

X.J.G. designed the methodology, conducted experiments, performed data analysis, and wrote and revised the manuscript. Z.S. and W.J.H. assisted in experimental execution and manuscript preparation. S.J.X. oversaw project administration. T.S.K. participated in manuscript review. Y.J. and Y.Z.P. contributed to data collection. Z.F. supervised study design, manuscript formatting, as well as data collection and annotation.

Funding

This work was financially supported by the National Natural Science Foundation of China (62401259), the Postdoctoral Fellowship Program of CPSF under Grant Number GZC20242228, the Fundamental Research Funds for the Central Universities (NZ2024036), and High Performance Computing Platform of Nanjing University of Aeronautics and Astronautics.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

This study has received approval from the Medical Ethics Committee of the Affiliated Eye Hospital of Nanjing Medical University and adhered to the principles outlined in the Declaration of Helsinki.

Additional information

Correspondence and requests for materials should be addressed to J.X. or F.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025